13.2 计数田鼠

工作纸

题目

STEP II 2000，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

step.maths.org/questions/00-s2-q13

提示

请先自行尝试解题。如果遇到困难，请逐一阅读提示，每读完一个便回到题目再试。标有停一停，试一试的提示是关键步骤。每个提示之后均留有空位供你演算。

> **想一想**　种群共有 N 只田鼠，其中 200 只带有记号。现随机抽取 200 只作为样本。恰好有 11 只带有记号的概率是多少？

阶段 1　似然（likelihood）

涵盖：概率 p_N

> **想一想**　在所有等可能的 200 只样本（sample）之中，从有标记的田鼠中选出 11 只，再从没有标记的田鼠中选出 189 只。然后把含有 N 的二项式系数（binomial coefficient）写成阶乘（factorial）的形式。

阶段 2 对整数求最大值

涵盖：求 p_N 的最大值

> **停一停，试一试** N 是整数，所以应比较 p_{N+1} 与 p_N（或 p_N 与 p_{N-1}）。大部分阶乘都会约去。这个比值何时大于 1？

阶段 3 第三个样本

涵盖：第三个样本

想一想 抽取第二个样本之后，有多少只田鼠带有标记？按照你对 N 的估计，有多少只没有标记？写出第三个 200 只的样本中恰好有 j 只有标记田鼠的概率。这些概率加起来等于甚么？

背后的数学

最大似然

最大似然原理由费雪（Ronald Fisher）在 1910 至 1920 年代发展出来，内容是：以使观察到的数据出现概率最大的那个值，作为未知参数的估计。它是至今应用最广的估计方法，由以曲线拟合数据，到训练机器学习中的神经网络，无一不用；神经网络的“损失函数”（loss function）通常就是负对数似然。费雪证明了，对于大样本，最大似然估计值近似服从正态分布（normal distribution），而且其精确程度不逊于任何合理的估计量（estimator），前提是模型具备某些良好性质，而这些性质对大多数日常模型都成立。不过也有重要的例外。在下一节（Section 13.3）中，未知量是可能出现的最大值，即分布的一个端点，在那里最大似然估计值的表现大不相同：它的误差并不近似服从正态分布，而且误差是以 $\frac{1}{n}$ 而非 $\frac{1}{\sqrt{n}}$ 的速度缩小。

这个估计有多精确？

图 13.2.1（见网上该节的完整版）显示似然曲线很宽阔：11 次重捕所提供的信息并不多。粗略计算一下：第二个样本中有记号田鼠的数目近似服从泊松分布（Poisson distribution），平均值为 $\frac{200\times 200}{N}$，所以其标准差（standard deviation）约为 $\sqrt{11}\approx 3.3$，相对误差约为 30%，而这个误差会传递到 N 的估计上。若要把 N 的估计误差控制在 10% 以内，生物学家需要大约 100 次重捕。统计学家还会修正这个简单估计的轻微偏差（bias），采用查普曼估计量（Chapman estimator）$\frac{(M+1)(C+1)}{R+1}-1$，其中 M 和 C 是两次捕捉的数量，R

是重捕到的数目。

拉普拉斯计数法国人口

拉普拉斯（Laplace）在 1802 年用了同样的想法，在没有人口普查的情况下估计法国的人口。他知道全国一年内登记的出生人数，并且就一批抽样的市镇而言，同时知道其人口和出生人数。他假设人口与出生人数之比在各地都相同，从而估计法国人口约为 28 百万，甚至给出了不确定程度的度量。后来这个方法在动物种群研究中被重新发现：C. G. J. Petersen 在 1890 年代于丹麦的峡湾为鲽鱼做记号，Frederick Lincoln 则在 1930 年代使用鸟环；这个方法至今仍常被称为林肯—彼得森（Lincoln–Petersen）估计量。

延伸练习

1 **一般公式。**若有 M 只动物做了记号，而第二个含 C 只动物的样本中有 R 只带有记号，证明 N 的最大似然估计值是 $\frac{MC}{R}$ 的整数部分。

2 **范德蒙恒等式。**考虑从 a 名女士和 b 名男士中选出 c 人组成委员会的方法数目，以此证明 $\sum_j \binom{a}{j}\binom{b}{c-j} = \binom{a+b}{c}$。

3 **掷硬币。**一枚硬币掷出正面的概率 p 未知，掷 n 次后观察到 h 次正面。证明 p 的最大似然估计值是 $\frac{h}{n}$。

4 **没有重捕。**如果第二个样本中没有任何带记号的田鼠，最大似然估计值会变成怎样？这说明了这个方法的甚么特点？

延伸阅读

- **David Spiegelhalter, *The Art of Statistics***（Pelican），谈估计与不确定性。
- **George Seber, *The Estimation of Animal Abundance***（Blackburn）。标记重捕法的标准参考书。
- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury）。数理统计的入门课本，内容包括最大似然。

完整的引导式解答见第 13.2 节：

https://beyondstep.keithmcnulty.org/zh-hans/ch13/voles.html

每星期一道 STEP 题目，完整讲解：The STEP Up，以电子邮件免费发送，订阅网址：keithmcnulty.substack.com